Intuitive Physics in Text-Conditional Image Generation

Eli Kramer (ekramer1@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge MA, 02139

Sarah Schwettmann* (schwett@mit.edu)
CSAIL, MIT
Cambridge MA, 02139

Pawan Sinha* (psinha@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge MA, 02139

* Indicates dual senior authorship

Abstract

Intuitive physical reasoning is one of the hallmarks of common-sense intelligence. We explore whether a contemporary machine learning model that can produce photo-realistic images demonstrates this aspect of common-sense visual intelligence. Specifically, we investigate whether DALL-E 2 (a state-of-the-art image generative model) has sufficient physical representations to generate plausible real-world images. We evaluate DALL-E's intuitive physics in four fundamental domains: stability, mass, refraction, and shadow. We find that DALL-E's creations are more congruent with human ratings in the domain of optics (refraction and shadow), than in physical dynamics (stability and mass).

Keywords: Intuitive physics; Image generation; Generative AI; Machine learning; Cognitive science

Introduction

The massive scaling of large pre-trained AI models has prompted researchers to claim that AGI–the holy grail of artificial intelligence–is within reach (Bubeck et al., 2023). Yet one recurring criticism of contemporary large machine learning models is that they lack common-sense intelligence (Lake et al., 2017; Marcus et al., 2019). It is generally believed that intuitive physics is a foundational component of common-sense intelligence (Zhu et al., 2020; Spelke et al., 1992).

Currently, intelligence benchmarks lag behind the rapid pace of AI model releases. Recent advances in image generation provide an opportunity to evaluate the physical reasoning abilities of large machine learning models. This paper develops a novel methodological approach to benchmarking the physical reasoning abilities of image generation models. We explore DALL-E 2, a diffusion model created by OpenAI in 2022 (Ramesh et al., 2022).

DALL-E was not explicitly programmed to have physical reasoning abilities, but the possibility that certain capabilities can spontaneously emerge is not without precedent in large pretrained models (Wei et al., 2022). Despite deficiencies in relational understanding (Conwell & Ullman, 2022) and geometric inconsistencies (Farid, 2022), DALL-E scores very highly on photo-realism and user preference (Ramesh et al., 2022). Critics have previously argued that pixel-level representations cannot model real-world physical dynamics (Ullman et al., 2018). We analyze whether large-scale text-image training implicitly equips DALL-E with knowledge grounded in real-world physics in four fundamental domains: stability, mass, refraction, and shadow.

Methods

We evaluate the physical realism of imagery generated by DALL-E compared to real-world photographs in four domains: mass, stability, refraction, and shadow. We use prompts that isolate physical dimensions and provide limited additional information (see Figures 1 and 2 for example stimuli). In Experiment 1 we generate a set of toppling tower stimuli with varying physical stability using the prompts "A(n) (un)stable tower of Jenga blocks." Experiment 2 evaluates DALL-E's representation of object mass. Heavy and light stimuli are generated using the prompts "A smooth lightweight (or heavy) red ball on a bed." The first two experiments consist of 90 images each (10 ground truth + 80 DALL-E). Experiments 3 (refraction) and 4 (shadow) use DALL-E's in-painting tool to alter real-world photographs. To investigate refraction, we created a dataset using a glass sphere that inverts and refracts the surrounding scene. We photographed the glass sphere in 40 different natural scenes and used the in-painting tool (Figure 2c) to erase the sphere's scene representation and reconstruct it with DALL-E using the prompt: "glass sphere." The refraction dataset contains 200 total images (40 ground truth + 160 DALL-E; we use all four DALL-E reconstructions per real-world image). We followed a similar process to create the shadow dataset, taking photographs of a single object on a desk with a directed light-source (from either the right or left side). We then erased evenly around the object to avoid biasing the position of the shadow (Figure 2f). The dataset

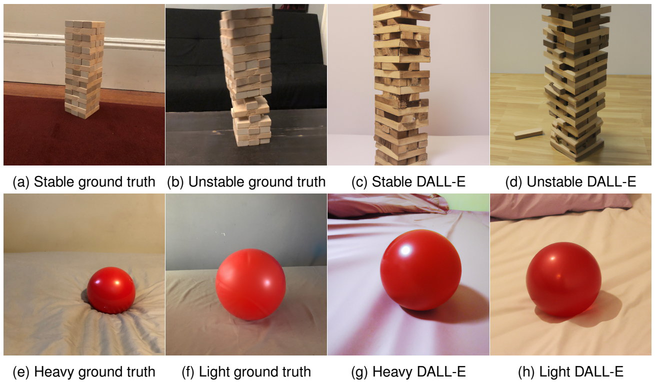
(a) Stable ground truth (b) Unstable ground truth (c) Stable DALL-E (d) Unstable DALL-E
(e) Heavy ground truth (f) Light ground truth (g) Heavy DALL-E (h) Light DALL-E

Figure 1: Example stimuli from mass and stability experiments, including ground truth and DALL-E generated images.

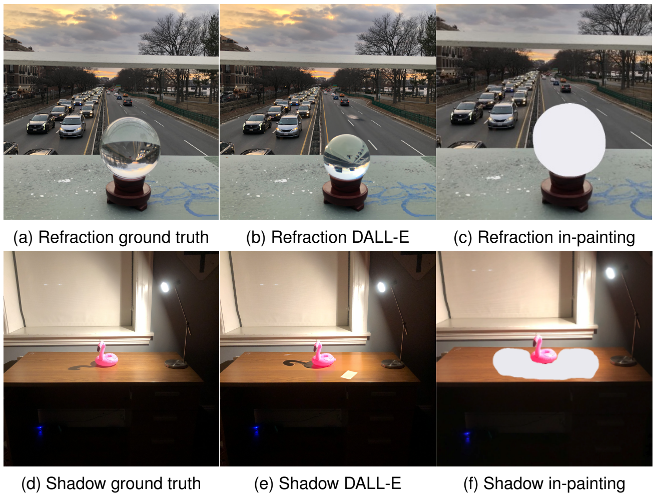
(a) Refraction ground truth (b) Refraction DALL-E (c) Refraction in-painting
(d) Shadow ground truth (e) Shadow DALL-E (f) Shadow in-painting

Figure 2: Example stimuli from refraction and shadow experiments, including ground truth DALL-E generated images, and in-painting technique.

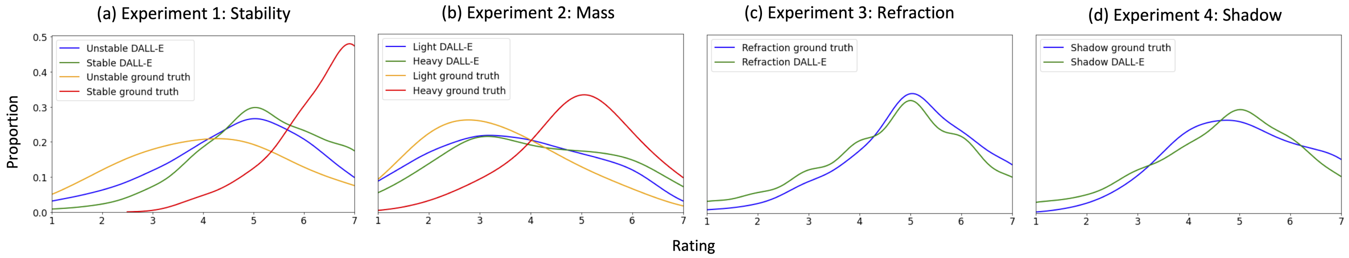

Figure 3: Distributions of human ratings for generated and ground truth stimuli in each experiment. The X-axis represents the human intuitive physics ratings from 1 to 7. The Y-axis is the proportion of observations made per image-group from 0 to 0.5. (Curves are smoothed for clarity using a bw_adjust = 1.5 in the Seaborn libarary's KDE plot)

contains 100 images (20 ground truth + 80 DALL-E). For all images (including the ground truth), we replace the DALL-E signature in the bottom right with an average of the colors directly surrounding the region.

DALL-E generated images are interleaved with ground truth images in four studies on Amazon Mechanical Turk. We select US-based participants with approval rate $>$ 95% for $>$100 HITs. Participants are instructed to rate a named physical characteristic in all images on a discrete scale from 1 to 7. Instructions include examples on both poles for all experiments.

Results

Figure 3 plots the distributions of human intuitive physical judgments on all generated and ground truth images for each experiment. In all analyses, we remove outlier participants whose mean ratings lie more than two standard deviations away from the mean across all participants (Miller, 1991). We note that more conservative thresholds (e.g. using three standard deviations) fail to remove any outliers; our approach excludes 18, 16, 38, and 16 participants who did not meet the criteria. The final sample sizes for the four experiments (mass, stability, refraction, shadow) were 270, 245, 558, and 311, with total judgements of 841, 805, 1713, and 815, respectively.

Experiment 1: Stability Participants rated real-world and generated Jenga towers on a stability scale ranging from 1 (very unstable) to 7 (very stable). We find no meaningful difference between the unstable and stable generated stimuli. The mean ratings for images generated with the stable and unstable prompts are 5.2 and 4.7, compared to 6.4 and 4.2 for ground-truth stimuli. Using the two-sample Kolmogorov–Smirnov (KS) test, we report a significant difference between the ratings on images generated with the stable vs. unstable prompt (0.12, $p = 0.003$), but the Jensen-Shannon divergence test reveals a divergence of only 0.086.

Experiment 2: Mass Participants rated real-world and generated images of a weighted ball on a bed on a scale ranging from 1 (very light) to 7 (very heavy). Using a KS test, we find no significant difference between the distributions for light (mean $= 3.6$) and heavy (mean $= 4.1$) generated stimuli (0.09, $p = 0.096$), suggesting DALL-E does not diferentially represent object mass. The Jensen-Shannon divergence between light and heavy distributions is 0.065. Ground truth heavy and light have mean ratings of 3.3 and 5.0.

Experiment 3: Refraction Next, we evaluate the refraction experiment, in which participants rated the physical validity of scene representations inside real-world (mean $= 5.1$) and generated glass spheres (mean$= 4.7$) on a scale from 1 (very incorrect) to 7 (very correct). We find no meaningful difference between the ground truth stimuli and DALL-E generated glass spheres. This suggests that DALL-E is capable of generating plausibly correct reconstructions. The KS test reports a significant difference between the two groups (0.11, $p = 0.002$) but with a low Jensen-Shannon divergence of 0.056.

Experiment 4: Shadow Finally, participants rated real-world (mean $= 5.1$) and generated (mean $= 4.7$) object shadows on a scale from 1 (very incorrect) to 7 (very correct). The KS test reveals no significant difference between ratings of ground truth and generated shadows (0.088, $p = 0.26$, Jensen-Shannon divergence $= 0.052$), suggesting DALL-E is adept at reconstructing shadows from scene illumination.

Discussion

We find that on average, human observers cannot distinguish between unstable vs. stable and heavy vs light DALL-E generated stimuli, whereas they are able to distinguish real-world variations along the same physical dimensions. In optics-based domains (refraction and shadow), however, we find that human observers do not meaningfully distinguish between real and DALL-E reconstructed versions of the same stimuli, suggesting that DALL-E is better equipped to generate plausible physical optics than variations in physical dynamics (stability and mass). While there is a statistically significant difference between the stable and unstable generated images, we hypothesize that this variation is not due to the stability of the structure itself but rather overt clues around the image (such as Jenga blocks on the floor, see Figure 2a). The Jensen-Shannon divergence test and visual inspection of the distribution plots corroborates this view.

Another potential explanation for convincing generated optics may be that human observers are less able to judge nuances of shadow and refraction (Nightingale et al., 2019; Ostrovksy et al., 2005). Future work will use sophisticated observers to discern which aspects of genuine optics appear DALL-E's creations. We can conclude that at the level of intuitive physics perceived by humans, there is a clear dichotomy between DALL-E's inability to generate plausible physical dynamics and its capacity to generate plausible optics.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Conwell, C., & Ullman, T. (2022). Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*.

Farid, H. (2022). Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Marcus, G., & Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Vintage.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, *43*(4), 907–912.

Nightingale, S. J., Wade, K. A., Farid, H., & Watson, D. G. (2019). Can people detect errors in shadows and reflections? *Attention, Perception, & Psychophysics*, *81*, 2917–2943.

Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies in scenes. *Perception*, *34*(11), 1301–1314.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological review*, *99*(4), 605.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive psychology*, *104*, 57–82.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... others (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.